Data Science e Tecnologie per le Basi di Dati
Homework 3

Sono date le relazioni seguenti (le chiavi primarie sono sottolineate):

```
IMPRESA-PULIZIE(Pid, Nome, Indirizzo, Città, Regione)
SERVIZI-OFFERTI(Pid, Sid)
SERVIZIO(Sid, NomeServizio, Categoria)
EDIFICIO(Eid, NomeEdificio, TipoEdificio, Indirizzo, Città , Regione)
SERVIZI-PULIZIA(Pid, Eid, Data, Sid, Costo, NumeroOre)
```

Si ipotizzino le seguenti cardinalità:

- card(IMPRESA-PULIZIE) = 10^4 tuple,
  valori distinti di Regione = 20
- card(SERVIZI-OFFERTI)= $2 \cdot 10^5$ tuple,
- card(SERVIZIO)= 100 tuple,
  valori distinti di Categoria = 10
- card(EDIFICIO)= $5 \cdot 10^7$ tuple,
  valori distinti di Città = 1000
  valori distinti di TipoEdificio = 10
- card(SERVIZI-PULIZIA)= 10^9 tuple,
  MIN(Data) = 1/1/2013, MAX(Data) = 31/12/2022

Inoltre si ipotizzi il seguente fattore di riduzione per le condizioni di group by:

- having COUNT(*)$\geq$1 $\simeq \frac{1}{2}$.
- having SUM(Costo)$\geq$1000 $\simeq \frac{1}{10}$.

Si consideri la seguente query SQL:

```
select Eid, SUM(Costo) as TotCost, SUM(NumeroOre) as TotOre
from SERVIZI-PULIZIA SP, EDIFICIO E
where SP.Data>=1/1/2022 and SP.Data<=31/12/2022
  and E.TipoEdificio <> 'Ufficio'
  and E.Città='Torino'
  and SP.Eid=E.Eid
  and SP.Sid IN ( select SO.Sid
                  from IMPRESA-PULIZIE IP, SERVIZIO S, SERVIZI-OFFERTI SO
                  where SO.Sid=S.Sid and SO.Pid=IP.Pid
                    and (Regione='Piemonte' or Regione='Liguria')
                    and Categoria='Interni'
                  group by SO.Sid
                  having COUNT(*)>=1)
group by SP.Eid
having SUM(Costo)>=1000
```

Homework tasks
Per l'interrogazione SQL

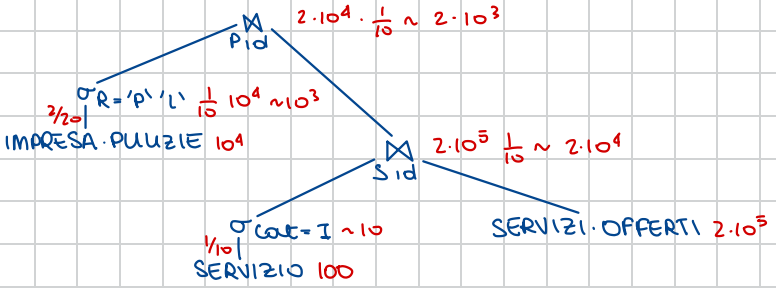
1. Si scriva l'espressione algebrica corrispondente, indicando le operazioni svolte, la cardinalità e la selettività di ogni operazione. Dove necessario, si ipotizzi la distribuzione dei dati. Discutere la possibilità di anticipare l'operatore GROUP BY.
2. Si scelgano le strutture fisiche accessorie per migliorare le prestazioni dell'interrogazione. Si motivi la scelta e si definisca il piano di esecuzione (ordine e tipo dei join, accesso alle tabelle e/o indici, etc.).

3 possibilità per la join interna:

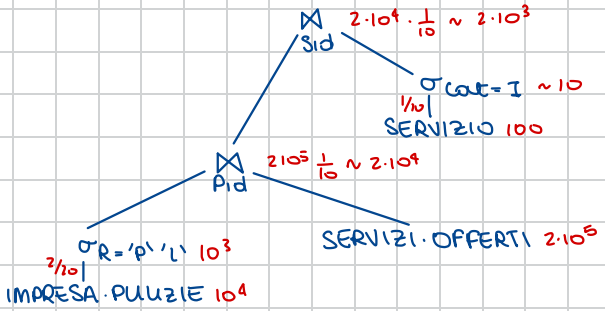
1. IMPRESA-PULIZIE $\bowtie_{Pid}$ (SERVIZIO $\bowtie_{Sid}$ SERVIZI-OFFERTI)
2. (IMPRESA-PULIZIE $\bowtie$ SERVIZIO) $\bowtie$ SERVIZI-OFFERTI
3. (IMPRESA-PULIZIE $\bowtie_{Pid}$ SERVIZI-OFFERTI) $\bowtie_{Sid}$ SERVIZIO)

↳ solo due join possibili: 1 e 3

①

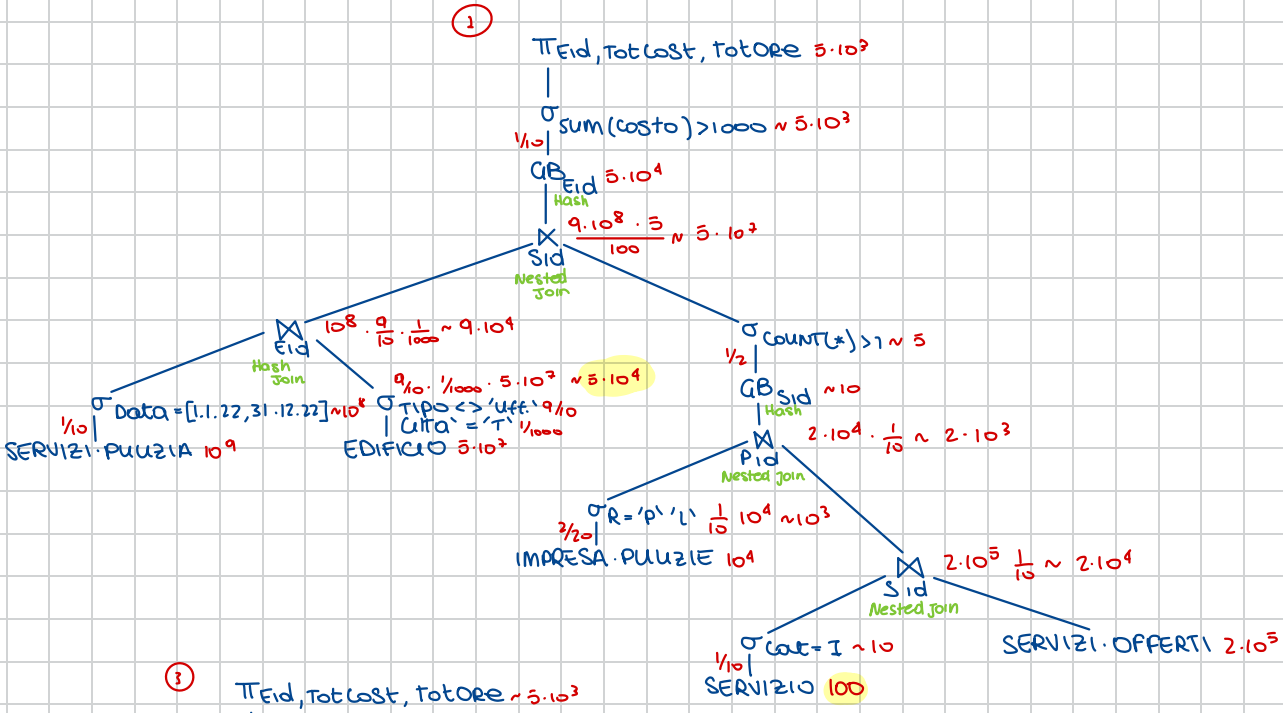


③

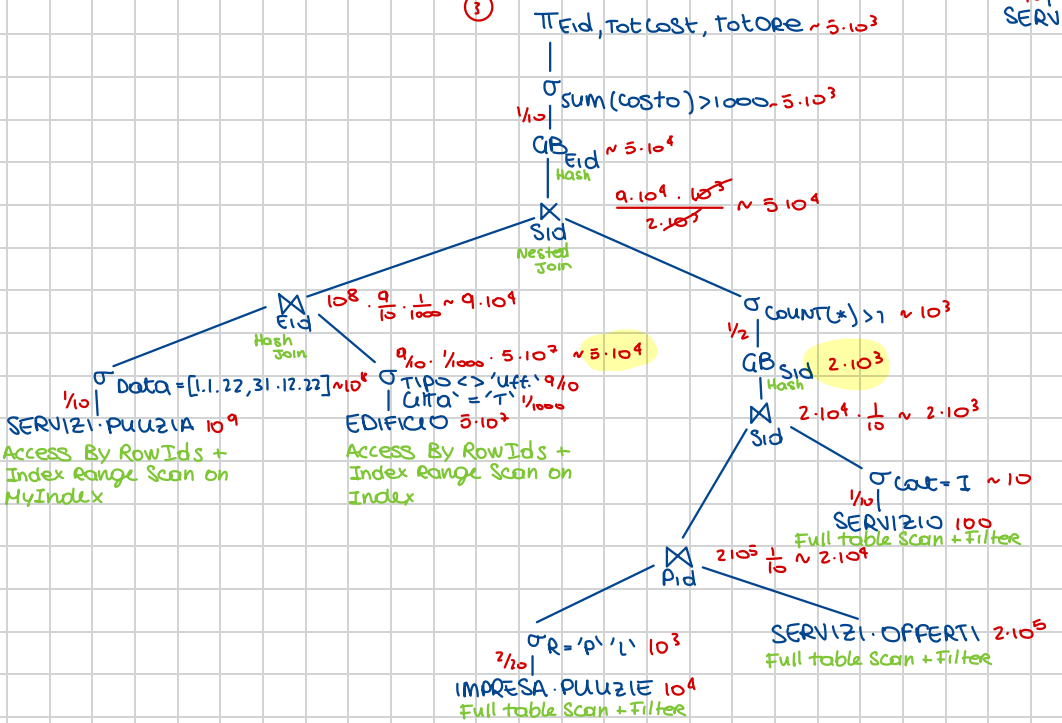


→ sono uguali ← scelgo la 3 per poter fare l'anticipazione della GB

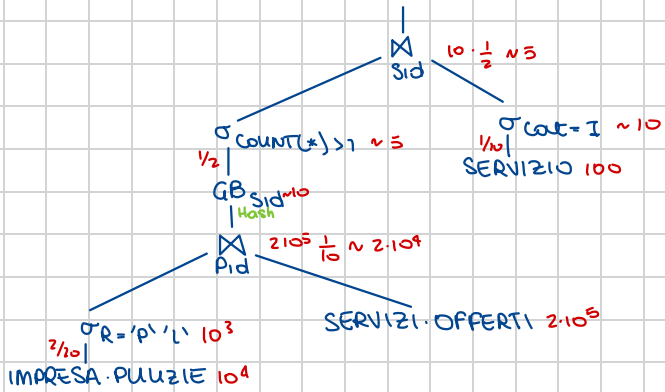
①



③



Nella query interna posso anticipare la GB_{Sid} sul ramo sinistro della $Join_{Sid}$



PER SERVIZI·PULIZIA creo un indice secondario sull'attributo Data di tipo Hash (uguaglianza), in quanto di grandi dimensioni e media selettività

CREATE INDEX MyIndex ON SP (Data);

PER EDIFICIO creo un indice secondario su Città di tipo B+ Tree (Range) in quanto di grandi dimensioni e alta selettività

CREATE INDEX Index ON E(Città)